Response to the comments of Xiaolong Fan (EGUSPHERE-2024-2496)

The Single-particle soot photometer (SP2) is a widely recognized instrument for quantifying the mixing state of black carbon (BC). However, deriving BC mixing state from SP2 measurements remains challenging. This study introduces a user-friendly SP2 inversion method based on machine learning. Notably, the machine learning approach does not merely replicate the results of physical inversion methods but also utilizes previously unexploited signals. It overcomes the low signal-to-noise ratio issue in input signal prevalent in conventional methods. This advancement will benefit the development of BC mixing state observations and radiative effect assessments. Overall, the manuscript is well-organized, and I recommend its publication after minor revisions.

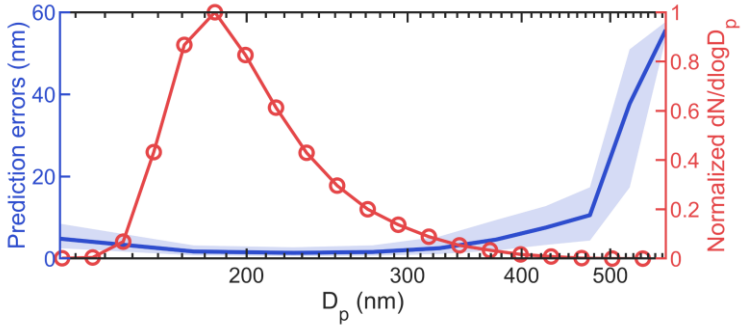
Response: We are grateful for the reviewer's valuable comments. We have carefully revised the manuscript by: (1) adding a detailed analysis of the model's performance across different particle diameter ranges, (2) including a comprehensive comparison of inversion results between training and testing sets in the Supplementary Information, and (3) enhancing the introduction section with a detailed rationale for selecting the LightGBM algorithm. Please find our responses (blue text) to the comments (black text) below.

1) There appears to be a correlation between the deviation of predicted values from the true values and particle size, as observed in Figure 3c. It would be beneficial to further characterize the relationship between prediction accuracy and particle diameter (D_p). This analysis could provide valuable insights into the model's performance across different particle size ranges and potentially identify any systematic biases or limitations in the prediction methodology.

Response: Thank you for pointing this out. We have added a comprehensive prediction error analysis for the D_p inversion model of internally mixed BC to further characterize the relationship between prediction accuracy and D_p. The added analysis can be found on Lines 310 to 323 in the revised manuscript and attached below:

*"**To comprehensively assess the model's performance across different particle size ranges, we further analyzed the prediction error distribution for D_p inversion model of internally mixed BC, as shown in Fig. 5. For particles smaller than 150 nm, the prediction errors average around 4 nm, primarily due to the low signal-to-noise ratio of their scattering signals, which introduces larger uncertainties in the LEO fitting process. The model exhibits optimal performance for particles between 150 nm and 300 nm, with an average prediction error of approximately 1.5 nm. Furthermore, based on the 25% and 75% percentiles of the error distribution, the model's prediction errors exhibit minimal fluctuation within this size range. However, prediction errors gradually increase with particle size, becoming particularly significant for particles larger than 480 nm. This trend can be attributed to occasional irregular signals at larger sizes, such as scattering or incandescence signals with abnormally broad peak widths. These signal irregularities pose challenges to the accurate characterization of particle physical properties,***

affecting both LEO fitting accuracy and ML model predictions, potentially leading to more pronounced discrepancies between the two methods. The number size distribution of internally mixed BC in the testing set indicates that most particles fall within the 150–300 nm range, where the model demonstrates highest accuracy. Although the prediction errors are relatively larger at both ends of the size distribution (< 150 nm and > 400 nm), the number of particles in these ranges is comparatively small, thus having limited impact on the overall performance of the model."



"***Figure 5. The prediction error distribution for D_p inversion model of internally mixed BC, and normalized number size distribution for D_p of internally mixed BC in the testing set. The solid lines in error distribution represent the median value, the upper and lower boundary of the is between the 25 % and 75 % quantiles.***"

2) Does the deviation between the predicted values and the true values refer to the test set, or does it also occur in the training set? What could be the underlying reasons for this? Please clarify.

Response: Thank you for your question. The observed deviation between predicted and true values exists in both training and testing sets. As shown in Figure R1, the coefficients of determination R^2 for the training and testing sets are 0.99 and 0.98, respectively. These high R^2 values indicate excellent model performance, with the close R^2 values demonstrating the model's strong predictive capability and good generalization performance.

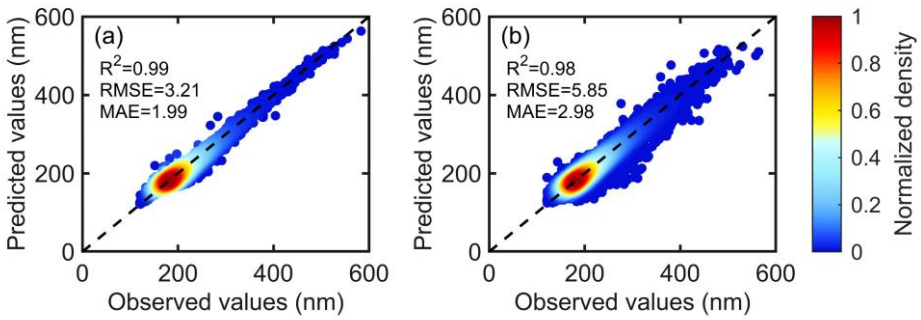


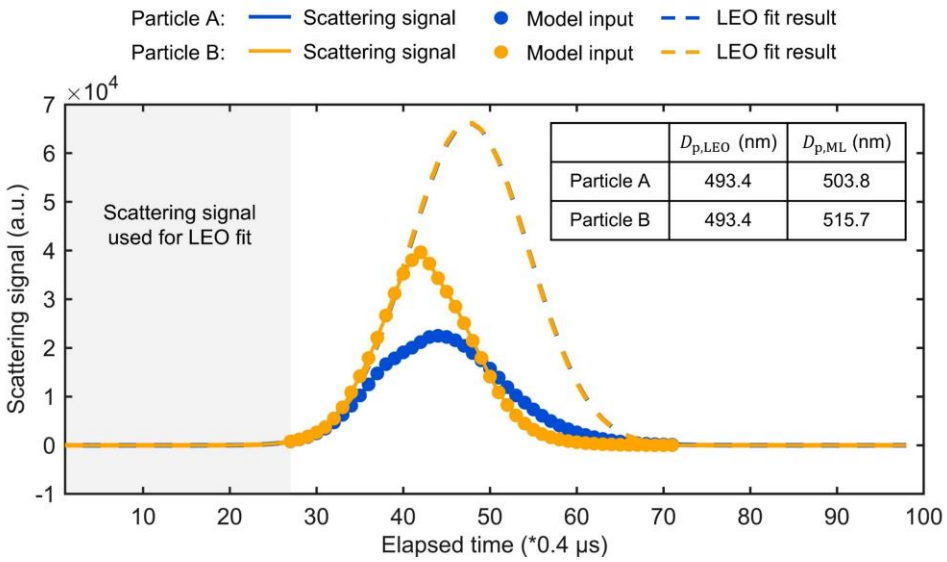
Figure R1. The D_p inversion results of internally mixed BC for both training set (a) and testing set (b).

The SP2 measurement process introduces intrinsic variability primarily through instrument background noise and measurement uncertainties. These factors contribute to inevitable errors and challenges in achieving high-precision measurements, even when employing advanced algorithms like LightGBM. Additionally, BC-containing particles exhibit complex signal characteristics marked by non-linear relationships, varying core-shell structures, and signal-to-noise ratio limitations, particularly for both ends of the particle size distribution, as we discussed in our response to Comment 1. These complexities affect both traditional physical inversion methods and machine learning predictions, leading to discrepancies between the "true" values (which have been renamed as "observed values" in the revised manuscript) obtained from physical inversion and the predicted values from machine learning models.

Furthermore, it's important to note that the LEO fitting method and ML method utilize different parts of the original signals, which can lead to discrepancies in D_p values. We conducted a comprehensive comparison between the LEO fitting method and the machine learning method, elaborating on the differences in signal utilization between the two approaches and their impact on inversion results. A detailed discussion can be found on Lines 395 to 402 of the revised manuscript. Moreover, the Fig. R1 and related content have been added as Fig. S3 to the Supplementary Information.

Lines 393 to 400:

"Figure 9 illustrates the LEO fitting results for two different BC-containing particles. Despite nearly identical leading-edge data resulting in similar Gaussian distributions and consequently the same D_p values through LEO fitting, the complete scattering signals of these particles exhibit significant differences. The ML model, by incorporating these distinctive signal features, can effectively capture these variations, leading to different D_p predictions. Moreover, the leading edge is traditionally defined as the signal from baseline-subtracted zero up to 5 % of the maximum laser intensity (Taylor et al., 2015). As shown in Fig. 9, this portion of the signal (in the grey-shaded area) is close to the baseline, making it particularly susceptible to noise interference. Compared to LEO fitting method, the ML model utilized a broad range of signals with a high signal-to-noise ratio, demonstrating enhanced noise resistance."



*"**Figure 9.** Comparison of the scattering signal used in the D_p inversion process for internally mixed BC and corresponding calculation results from both the LEO fitting and the ML methods. The solid line represents the scattering signal obtained by SP2, and the part marked with solid dots is the scattering signal input to the ML model. The gray shaded area shows the leading-edge data used in the LEO fitting process, and the dashed line represents the scattering signal of the original particle reconstructed by LEO fitting."*

3) Does the inversion of D_c and D_p in BC-containing particles utilize multiple outputs from the same trained model, or from different models? Additionally, does D_c influence the inversion of D_p?

Response: Thank you for your comment. This study employs different inversion models for D_p and D_c of BC-containing particles. Although both models are founded on the LightGBM algorithm, they are developed independently due to their distinct feature data and target variables. The D_c inversion model uses only incandescence signals as features, while the D_p inversion model incorporates both incandescence and scattering signals. Consequently, two separate models with different hyperparameters are required to establish unique mappings between their respective input signals and particle characteristics.

For internally mixed BC, D_c influences the inversion of D_p. In the traditional physical inversion method, we first obtain the peak height of the reconstructed scattering signal through LEO fitting, and then derive the scattering cross-section. This scattering cross-section is determined by both the BC core and coating material. Consequently, D_c needs to be integrated with Mie theory to accurately estimate D_p. A detailed introduction of this method is provided in the "Construction of label dataset" section of the revised manuscript.

In our machine learning method, we similarly consider the influence of D_c. The

characteristics of the BC core in internally mixed BC are reflected in the incandescence signal. Therefore, when constructing the D_p inversion model for internally mixed BC, we incorporate both scattering signals and incandescence signals as feature data. This methodology allows us to capture the influence of both the BC core and the coating on the overall particle size, enabling a more accurate prediction of D_p for BC-containing particles.

Lines 208 to 215:

"As the evaporation of the particle, the scattering signal deviates from a Gaussian distribution, making it inappropriate to directly use the scattering amplitude to calculate D_p. ***To properly size these particles, the LEO fitting method is employed to reconstruct the Gaussian signal. As described in Sect. 3.3, the zero-crossing point in the TEAPD signal can serve as a position reference for particles in the SP2. Moreover, the position difference between the zero-crossing point and the peak laser intensity remains constant during measurements. The width of the laser intensity distribution and the position of peak laser intensity relative to the zero-crossing point, both determined by Gaussian fitting of numerous unsaturated pure scattering particles, are used to constrain the LEO fitting, leaving the fitting amplitude as the only free parameter. Using leading-edge data from the signal onset to 5% of the maximum laser intensity for LEO fitting, can obtain the reconstructed scattering amplitude and further convert it to particle scattering cross-section.*** *The D_p of internally mixed BC can be derived by inputting the LEO-fitted scattering cross-section, BC core diameter, and the corresponding refractive indices of the core and coating into the Mie calculation model (Laborde et al., 2012; Liu et al., 2014; Schwarz et al., 2008; Taylor et al., 2015)."*

4) Could you elaborate on the rationale behind selecting LightGBM over alternative models?

Response: Thank you for your suggestion. We have provided a more comprehensive rationale for selecting LightGBM over alternative models in the introduction of the revised manuscript.

The relevant amendments are detailed on Lines 55 to 66:

"As an alternative, data-driven models such as machine learning (ML) can provide a good supplement to physical process-based models. ML can efficiently capture the nonlinear relationship between inputs and outputs, and has found widespread application in various fields (Carleo et al., 2019; Jordan and Mitchell, 2015; Liakos et al., 2018; Tarca et al., 2007). ***In recent years, tree-based machine learning models have gained considerable popularity due to their extremely high computational speed, satisfactory accuracy, and interpretability (Keller and Evans, 2019; Li et al., 2022; Wei et al., 2021; Yang et al., 2020). Among these, the Light Gradient Boosting Machine (LightGBM) has shown particularly outstanding performance. As a novel distributed gradient boosting framework based on decision tree algorithms, LightGBM can extract information from data more effectively than traditional tree models, excelling in handling complex non-linear relationships and high-dimensional features (Ke et al., 2017; Liu et al., 2024; Zhong et al., 2021). It employs innovative techniques such as gradient-based one-side sampling (GOSS) and exclusive***

feature bundling (EFB), which significantly improve computational efficiency while maintaining high predictive performance (Ke et al., 2017; Sun et al., 2020). Furthermore, different from some black-box models, LightGBM maintains the interpretability characteristic of tree-based models (Gan et al., 2021; Zhang et al., 2019), which can provide decision path analysis, allowing for deeper insights into the decision-making process. Considering these advantages, LightGBM can be an ideal tool for analyzing large SP2 datasets and inverting BC mixing states."

Reference

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L.: Machine learning and the physical sciences, Rev. Mod. Phys., 91, 045002, https://doi.org/10.1103/RevModPhys.91.045002, 2019.

Gan, M., Pan, S., Chen, Y., Cheng, C., Pan, H., and Zhu, X.: Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia River, JMSE, 9, 496, https://doi.org/10.3390/jmse9050496, 2021.

Jordan, M. I. and Mitchell, T. M.: Machine learning: Trends, perspectives, and prospects, Science, 349, 255–260, https://doi.org/10.1126/science.aaa8415, 2015.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, 30, 2017.

Keller, C. A. and Evans, M. J.: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10, Geoscientific Model Development, 12, 1209–1225, https://doi.org/10.5194/gmd-12-1209-2019, 2019.

Laborde, M., Mertes, P., Zieger, P., Dommen, J., Baltensperger, U., and Gysel, M.: Sensitivity of the Single Particle Soot Photometer to different black carbon types, Atmos. Meas. Tech., 5, 1031–1043, https://doi.org/10.5194/amt-5-1031-2012, 2012.

Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., and Geng, Y.: Application of XGBoost algorithm in the optimization of pollutant concentration, Atmospheric Research, 276, 106238, https://doi.org/10.1016/j.atmosres.2022.106238, 2022.

Liakos, K., Busato, P., Moshou, D., Pearson, S., and Bochtis, D.: Machine Learning in Agriculture: A Review, Sensors, 18, 2674, https://doi.org/10.3390/s18082674, 2018.

Liu, D., Allan, J. D., Young, D. E., Coe, H., Beddows, D., Fleming, Z. L., Flynn, M. J., Gallagher, M. W., Harrison, R. M., and Lee, J.: Size distribution, mixing state and source apportionment of black carbon aerosol in London during wintertime, Atmospheric Chemistry and Physics, 14, 10061–10084, 2014.

Liu, Z.-H., Weng, S.-S., Zeng, Z.-L., Ding, M.-H., Wang, Y.-Q., and Liang, Z.: Hourly land surface temperature retrieval over the Tibetan Plateau using Geo-LightGBM framework: Fusion of Himawari-8 satellite, ERA5 and site observations, Advances in Climate Change Research, 15, 623–635, https://doi.org/10.1016/j.accre.2024.06.007, 2024.

Schwarz, J. P., Spackman, J. R., Fahey, D. W., Gao, R. S., Lohmann, U., Stier, P., Watts, L. A., Thomson, D. S., Lack, D. A., Pfister, L., Mahoney, M. J., Baumgardner, D., Wilson, J. C., and Reeves, J. M.: Coatings and their enhancement of black carbon light

absorption in the tropical atmosphere, J. Geophys. Res., 113, 2007JD009042, https://doi.org/10.1029/2007JD009042, 2008.

Sun, X., Liu, M., and Sima, Z.: A novel cryptocurrency price trend forecasting model based on LightGBM, Finance Research Letters, 32, 101084, https://doi.org/10.1016/j.frl.2018.12.032, 2020.

Tarca, A. L., Carey, V. J., Chen, X., Romero, R., and Drăghici, S.: Machine Learning and Its Applications to Biology, PLoS Comput Biol, 3, e116, https://doi.org/10.1371/journal.pcbi.0030116, 2007.

Taylor, J. W., Allan, J. D., Liu, D., Flynn, M., Weber, R., Zhang, X., Lefer, B. L., Grossberg, N., Flynn, J., and Coe, H.: Assessment of the sensitivity of core / shell parameters derived using the single-particle soot photometer to density and refractive index, Atmos. Meas. Tech., 8, 1701–1718, https://doi.org/10.5194/amt-8-1701-2015, 2015.

Wei, J., Li, Z., Pinker, R. T., Wang, J., Sun, L., Xue, W., Li, R., and Cribb, M.: Himawari-8-derived diurnal variations in ground-level $PM_{2.5}$ pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM), Atmospheric Chemistry and Physics, 21, 7863–7880, https://doi.org/10.5194/acp-21-7863-2021, 2021.

Yang, L., Xu, H., and Yu, S.: Estimating PM2.5 concentrations in Yangtze River Delta region of China using random forest model and the Top-of-Atmosphere reflectance, Journal of Environmental Management, 272, 111061, https://doi.org/10.1016/j.jenvman.2020.111061, 2020.

Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., Wang, Q., and Huang, L.: A Predictive Data Feature Exploration-Based Air Quality Prediction Approach, IEEE Access, 7, 30732–30743, https://doi.org/10.1109/ACCESS.2019.2897754, 2019.

Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., and Zhang, W.: Robust prediction of hourly PM2.5 from meteorological data using LightGBM, National Science Review, 8, nwaa307, https://doi.org/10.1093/nsr/nwaa307, 2021.